SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1b. RESTRICTIVE MARKINGS

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release;
distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

D Cg

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TR- 89-0068

6a. NAME OF PERFORMING ORGANIZATION
Louisiana State University

6b. OFFICE SYMBOL (If applicable)

7a. NAME OF MONITORING ORGANIZATION
AFOSR/NM

6c. ADDRESS (City, State, and ZIP Code)
Computer Science Department
Baton Rouge, Louisiana 70803

7b. ADDRESS (City, State, and ZIP Code)
AFOSR/NM
Bldg 410
Bolling AFB DC 20332-6448

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
AFOSR

8b. OFFICE SYMBOL (If applicable)
NM

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
AFOSR-87-0160

8c. ADDRESS (City, State, and ZIP Code)
AFOSR/NM
Bldg 410
Bolling AFB DC 20332-6448

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
61102F	2304	A2	

11. TITLE (Include Security Classification)
Parametric Analysis of Queueing Networks with Blocking

12. PERSONAL AUTHOR(S)
Dr. Ian F. Akyildiz

13a. TYPE OF REPORT
Final

13b. TIME COVERED
FROM 4/1/87 TO 8/21/87

14. DATE OF REPORT (Year, Month, Day)
5/11/88

15. PAGE COUNT
21

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Queueing networks with blocking have experienced a dramatic increase in their importance regarding performance evaluation of computer systems and communication networks. Parametric Analysis is very interesting for cases in which only one station (e.g., a CPU) in a queueing network model is to be analyzed under various system washload. In order to execute parametric analysis of queueing networks with blocking the problem "Computation of the Throughput Values of the Finite Capacity Subsystem" is solved. The accuracy of the method has been validated by simulation of several test cases.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
□ UNCLASSIFIED/UNLIMITED □ SAME AS RPT. □ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclass

22a. NAME OF RESPONSIBLE INDIVIDUAL
~~Maj. John P. Thomas~~ Dr. Abraham Waxman

22b. TELEPHONE (Include Area Code)
-767-5027

22c. OFFICE SYMBOL
NM

83 APR edition may be used until exhausted.
All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

Final Report on the Project

"PARAMETRIC ANALYSIS OF QUEUEING NETWORKS WITH BLOCKING"

AFOSR-87-0160

I. F. Akyildiz
Computer Science Department
Louisiana State University
Baton Rouge, LA 70803

1. In the first step of the project we had to verify the suggested algorithm, (Throughput Analysis of Blocking Networks), described in the proposal on pages 7, 8 and 9, to be accurate. We studied several test examples and compared our results with simulaqtion obtained by IBM-RESQ simulation package. The study showed that the algorithm is very accurate. The results obtained showed deviations from the simulation counterparts on the average 4%. The paper, [1], where this algorithm is described in detail has been accepted by *IEEE Transactions on Computers*" in October 1987. The paper will appear in November 1988 or March 1989 issue.

2. While we were working on the validation of the throughput algorithm mentioned above, we developed another new throughput algorithm for blocking queueing networks having stations with general service time distributions and FCFS scheduling disciplines. The research results obtained in this part are so significant that the paper, [2], was accepted into "Performance 87" Conference where the acceptance rate was only 25%. We attended the conference in Brussels/Belgium in December 1987 and presented the paper.

3. The problem of determining the capacity of the composite station, section 3.2. of the proposal, has also been attacked and solved partially. Some ideas have been described in [3] where a computer network with local and global window flow control is analyzed. We suggest some formulas for the computation of the capacity of the composite (flow-equivalent) station. However, we realized that the formulas do not perform very accurate for all cases. There is still some need for

investigation of this part. The paper [3] was accepted to the INFOCOM 88 Conference. We attended the conference in New Orleans in March 1988 and presented the paper.

There is another type of blocking which is known as "*Rejection Blocking*" in the literature. In this case the blocking occurs when a job completing service at station i attempts to join destination station j. If station j is full at that moment, the job is rejected. The rejected job goes back to the server of the station i and receives another round of service. This is repeated until some job completes a service at station j and a place becomes available. This blocking type has been used to model systems such as production systems and telecommunication systems. In particular, in token ring networks, the station which has the token, may transmit. In case of nonsuccessful transmission the token comes back to the station which retries the transmission.

Within this project we also attacked queueing networks with this type of blocking and obtained the results described below:

1. In [4] we obtained an exact product form solution for equilibrium state probabilities for single class closed rejection blocking networks which have reversible routing. An algorithm is given for computation of performance measures. For nonreversible networks we found out that the state space of a blocking queueing network is isomorphic to the state space of a nonblocking network under a particular condition. This paper was accepted into the GI/NTG Conference. We attended the conference in Erlangen/West Germany in September 1987 and presented the paper.

2. In [5] we investigated open, mixed and closed queueing networks with multiple job classes, reversible routing and rejection blocking. Jobs could change class membership and load dependent general service time distributions were allowed. We prove that the equilibrium state probabilities have product form. The solution implies insensitivity in this kind of blocking networks, i. e. the distribution of the jobs in equilibrium, irrespective of their remaining service times. This work is also presented at the Conference "Analysis and Control of Large Scale Stochastic Systems", Chapel Hill, NC, May 23-25, 1988.

2. In [6] we studied the same model as in [5]. Using the product form of the equilibrium state distribution obtained in [5], we derive exact algorithms to compute performance measures, such as mean number of jobs and throughputs.

3. State-dependent routing is a very important issue in computer systems and communication networks. For particular infinite capacity networks there exist solutions in the literature. However, considering the state dependent routing with the blocking phenomenon in queueing network models, makes the analysis more complex. In [7] we attacked this problem and solved it for central server models with multiple job classes. Using the concept of job local balance, we prove that the equilibrium state probabilities of these networks take a modified product form solution. We also develop an algorithm for the computation of performance measures, like throughputs and the mean number of jobs, is given.

References

1. I. F. Akyildiz, "Product Form Approximations for Queueing Networks with Multiple Servers and Blocking", *to appear in IEEE Transactions on Computers,* (accepted in October 1987; to be published in November 1988 or March 1989).

2. I. F. Akyildiz, "General Closed Queueing Networks with Blocking", *Proc. Performance 87 Conference,* North Holland Publishing Co., editors P. J. Courtois and G. Latouche, December 1987, pp. 283-303

3. I. F. Akyildiz, "Performance Analysis of Computer and Communication Networks with Local and Global Window Flow Control", *Proc. of INFOCOM Conference,* New Orleans, LA, March 1988, pp. 401-411

4. I. F. Akyildiz, "Analysis of Reversible and Nonreversible Queueing Networks with Rejection Blocking", *Proc. of the GI/NTG Conference on Measurement, Modeling and Evaluation of Computer Systems,* Springer Verlag, editors U. Herzog and M. Paterok, September 1987, pp. 150-163

5. I. F. Akyildiz and H. von Brand, "Exact Solutions for Open, Closed and Mixed Queueing Networks with Rejection Blocking", *to appear in Theoretical Computer Science Journal,* North Holland Publ. Co.

6. I. F. Akyildiz and H. von Brand, "Computation of Performance Measures for Queueing Networks with Reversible Routing and Rejection Blocking", *to appear in ACTA Informatica Journal,* Springer Verlag.

7. I. F. Akyildiz and H. von Brand, "Central Server Models with Multiple Job Classes, State-Dependent Routing and Rejection Blocking", *Submitted for publication.*

Parametric Analysis of Queueing Networks with Blocking

I. F. Akyildiz

Department of Computer Science
Louisiana State University
Baton Rouge, Louisiana 70803
U. S. A.

1. Introduction

Queueing networks have a great popularity as models of computer systems since the early seventies, because they allow the modeling of multiple independent resources such as CPU's and I/O devices and the sequential use of these resources by different jobs. The basic results of queueing network theory were given by Jackson, Gordon/Newell [JACK63, GORD67a] where they showed that the solution of open and closed queueing networks with single job class, exponentially distributed arrival and service times, First-Come-First-Served queueing disciplines at all stations have a **product form**. This product form implies that the equilibrium state probabilities consist of factors representing the states of the individual stations. As a result the individual stations behave as if they were separate queueing systems. Baskett, Chandy, Muntz and Palacios [BCMP75] extended the results of [JACK63, GORD67a] to obtain product form solutions for open, closed and mixed queueing networks with different job classes, nonexponential service time distributions and different queueing disciplines such as First-Come-First-Served (FCFS), Processor Sharing (PS), Last Come First Served Preemptive Resume (LCFS-PR).

Product form queueing networks (also known as BCMP or separable networks) have proved invaluable in modeling a variety of computer and communication systems. They are flexible enough to represent adequately many of the features arising in such applications. They have not, however, been able to provide much insight into the phenomenon of **blocking**, because all algorithms for product form networks are based on the assumption that the stations have infinite capacities. If the stations have finite capacities, blocking can occur in the network.

Various types of blocking have been considered in the literature so far. These blocking types arose out of various studies of real life systems. We classify the blocking networks as "Classical Blocking" and "Rejection

Blocking" networks. In the first case, classical blocking, blocking occurs when a job completing service at station *i* cannot proceed to station *j* because station *j* is full. The job is forced to wait in station *i*'s server until it is allowed to enter the destination station *j*. Station *i*'s server stops processing until station *j* releases a job. This blocking type has been used to model systems such as production systems and disk I/O subsystems. In the second case, rejection blocking, blocking occurs when a job completing service at station *i* attempts to join destination station *j*. If station *j* is full at that moment, the job is refused. The rejected job goes with a certain probability (which we will call the rejection probability) back to station *i*'s server and receives a new service. This is repeated until some job completes a service at station *j* and a place becomes available. This blocking type has been used to model systems such as production systems and telecommunication systems.

In recent years there has been a growing interest in the development of computational methods to analyze queueing networks with blocking. The interest developed primarily from the realization that these models are useful in the study of system behavior of computers and communication networks, in addition to providing detailed descriptions of several computer-related applications.

Most of the previous work is based on investigating **"Rejection Blocking"** in both open queueing networks [KONH76,77] and closed queueing networks [BALS83, GORD67b, HORD81, PITT79, SURI84]. Konheim/Reiser [KONH76,77] propose an algorithm for the solution of an open system consisting of two single server stations with exponential service time distributions. It also permits a feedback from the second station to the first station's queue. Pittel [PITT79], Hordijk/Van Dijk [HORD81] and Balsamo/Iazeoalla [BALS83] have shown that the equilibrium state probability distribution has a product form, given that the **"reversibility"** condition holds in closed queueing networks with rejection blocking.

The Suri/Diehl [SURI84] study examined closed tandem queueing networks with finite station capacities in which the first queue has a capacity larger than the number of jobs in the system. By application of Norton's Theorem [CHAN75], they reduce (N-1) stations to a single station with a variable size queue capacity and obtain a two-station network that is easy to analyze. An approximation algorithm is derived for the mean sojourn time of a job, assuming exponentially distributed service times. The main disadvantages to this technique were that validation tests were restricted to networks with very small populations and their algorithm is restricted only to serially switched stations.

"Classical Blocking" networks have also been investigated extensively in recent years [AKYL85a,b,c,d, PERR81, PERR86, TAKA80]. In [AKYL85a] we studied two-station closed queueing networks with classical blocking and multiple server stations. We have shown that the equilibrium state probability distributions of such blocking systems are identical to those of a two-station closed queueing network without blocking. In [AKYL85b] we show that the throughput of a blocking network with K total number of jobs is approximately equal to the throughput of a nonblocking network with an appropriate total number of jobs $\hat{K}$, which can be easily calculated. In [AKYL85c] we introduce an approximation algorithm for obtaining the throughput and mean queue length of closed exponential queueing networks with blocking. In [AKYL85d] we extend the well-known mean value analysis algorithm [REIS80] to single server queueing networks with blocking. The approximation is based on the modification of mean residence times due to the blocking events that occur in the network.

Takahashi, Miyahara and Hasegawa [TAKA80] developed a method for approximate analysis of open queueing networks with classical blocking. Each station is treated as an M/M/1 finite capacity queueing system whose arrival rate and mean service time are expressed in terms of the blocking probabilities. These probabilities are in turn expressed in terms of the arrival rates and mean service times yielding a set of N simultaneous non-linear equations whose solution yields an approximation for blocking probabilities. Approximations for other performance measures can be obtained from these probabilities. However, only a very limited accuracy assessment was performed.

Perros [PERR81] considered a general class of open exponential queue networks consisting of n $(n \geq 2)$ queues in parallel being serviced by servers who form a hierarchical structure. Blocking of a server occurs each time the server completes a service. The server remains blocked until its blocking unit departs from the network having received service by all the other servers to which this server is linked. Approximate and exact results for the utilization of a station were obtained. Perros/Altiok [PERR86] analyze open queueing networks with exponentially distributed service times. Their work is based on communication networks in which each station is serially switched. A Cox distribution for each station is developed in which the second phase represents the blocking phase of the corresponding station. Blocking probabilities are determined using an iterative formula.

Several other investigators in recent years have published results on queueing networks with rejection as well as classical blocking. An excellent bibliography concerning queueing network models with blocking is given by

Perros [PERR84]. Our literature review suggests that existing or proposed methods either contain disadvantages (e.g., long run times and/or memory space) and/or restrictions (only two-station or tandem network solutions) or provide approximate results which differ widely from the exact values. The blocking network models which have been investigated so far have the following additional limitations:

i) All service time distributions are exponential.

ii) The queueing discipline at each station is basically FCFS.

iii) All stations may have single (load independent) servers.

In this proposal we will attack these limitations and propose new solutions for closed queueing networks with blocking.

2. Model Assumptions

We consider closed queueing networks with N stations and K total jobs. Each station in the network may have the following four station types:

- **Type 1a.** [$M / M / 1 - FCFS$]
- **Type 1b.** [$M / M / ld - FCFS$] (ld: load-dependent server; allows multiple servers)
- **Type 2.** [$\bullet / G / 1 - PS$]
- **Type 3.** [$\bullet / G / IS - .$]
- **Type 4.** [$\bullet / G / LCFS\text{-}PR$]

Each station of Type 1a or Type 1b has exponential service time distribution and of Type 2,3,4 general service time distribution with mean values $1/\mu_i$ for $i = 1,...,N$. The service rate of Type 1b station is load dependent $\mu_i(\underline{k})$. Note that four types of stations are motivated by some practical considerations. Type 1 stations are useful in many instances (secondary memory units, input-output devices, etc.). Type 2 stations are, in many cases, a reasonable representation of the CPU allocated in quanta; the processor sharing discipline is an idealization with a quantum of "infinitesimally small" (in fact, zero) duration and no overhead associated with switching from one job to the other. A Type 3 station represents well terminals in a time-sharing system. Type 4 stations can be used to represent stacks in data structure models.

Each station also has a fixed finite capacity M_i where $M_i = (\mathit{queue\ capacity} + 1)$, (for $i = 1,2,...,N$). Cases in which the stations can have infinite capacity are also allowed. ($M_i = \propto$), (for some $i = 1 , 2 , .., N$). Any station whose capacity exceeds the total number of jobs in the network can be considered to have infinite capacity. A job which is serviced by the i-th station proceeds to the j-th station with probability p_{ij}, (for $i , j = 1 , 2 , \cdots , N$), if the j-th station is not full. That is, if the number of jobs in the j-th station, k_j, is less or equal to M_j for j=1,2,...,N. Otherwise, the job is blocked in the i-th station until a job in the j-th station has completed its servicing and a place becomes available.

Furthermore, we assume that

$$K < \sum_{i=1}^{N} M_i \tag{1}$$

which implies that the total number of jobs K in the network may not exceed the total station capacity of the entire network.

One of the most important problems to realize regarding blocking queueing networks is that finite station capacities and blocking can introduce the problem of system deadlock. Deadlock may occur if a job which has finished its service at station i's server wants join station j, whose capacity is full. That job is blocked in station i. Another job which has finished its service at j-th station now wants to proceed to the i-th station, whose capacity is also full. It blocks station j. Both jobs are waiting for each other. As a result a deadlock situation arises. In [AKYL86a] we have demonstrated the necessary conditions for a closed queueing network with single job class to be deadlock free. The following assumption states that a closed queueing network containing finite station capacities is **deadlock free** if and only if for each cycle C in the network the following condition holds:

$$K < \sum_{j \in C} M_j \tag{2}$$

Simply stated, the total number jobs in the network must be smaller than the sum of station capacities in each cycle. Since tandem queueing networks have only one cycle, this condition, equation (2), corresponds to equation (1). Equation (1) is a sufficient condition for tandem networks to be deadlock free.

With these assumptions we obtain the queueing network model, classified as classical blocking which will be the object of our investigation.

3. Norton's Theorem Application on Blocking Queueing Networks (PROPOSED CONCEPT I)

The parametric analysis is based on an application of Norton's Theorem from electrical circuit theory to queueing networks. Chandy, Herzog and Woo [CHAN75] showed that Norton's Theorem provides an exact analysis of queueing networks, if such networks have a product form solution. We explain this concept by considering a closed queueing network model with K jobs and N = 12 stations shown in Figure 1.

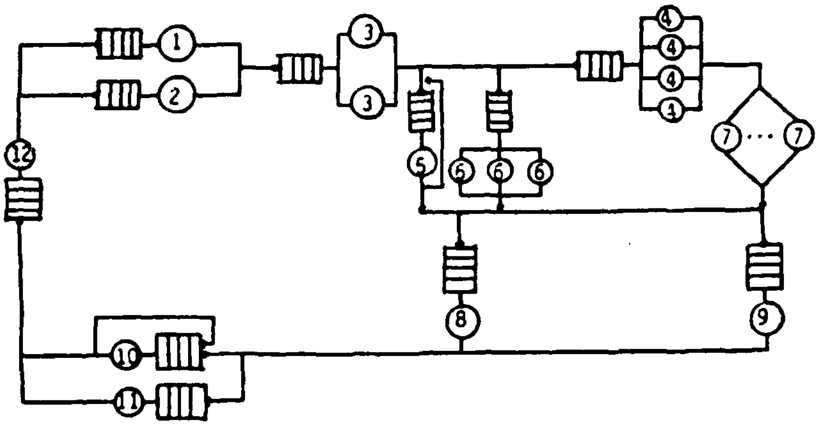
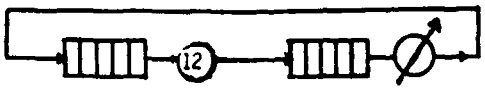


Figure 1.

With this queueing network model we can construct an equivalent network in which we arbitrarily select a station N, and replace the other $(N - 1)$ stations by a single station, called the composite (flow-equivalent) station as shown in Figure 2.



"composite station"

Figure 2.

Let $\mu_c(k)$ be the composite mean service rate, where k is the number of jobs at this composite station. These composite mean service rates $\mu_c(k)$, (for $k = 1,...,K$), are determined by analyzing a modified version of the given network, in which the selected station N has been shorted as shown in Figure 3.

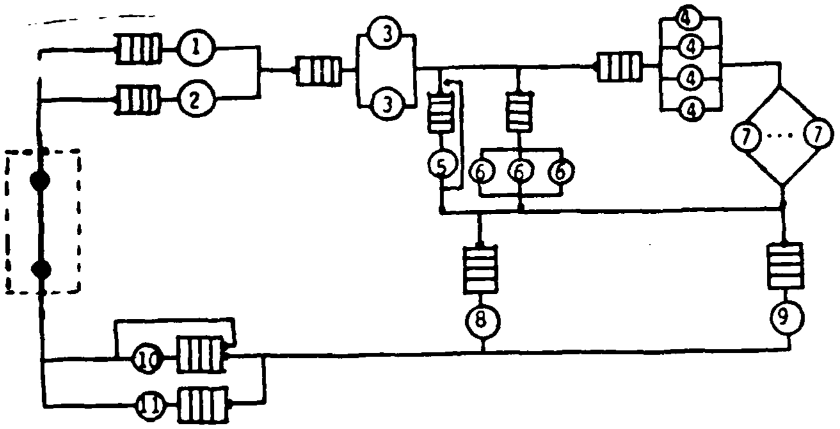


Figure 3.

The mean service time of that station $1/\mu_N$ is set equal to zero and the throughputs $\lambda(k)$, of the shorted system for all jobs $k=1,2,..,K$ are calculated. These computed throughputs through the shorted network $\lambda(k)$, are set equal to the composite mean service rates $\mu_c(k)$. The solutions of the network consisting of the selected and composite stations are identical to those of the originally given network model.

The parametric analysis of Chandy, Herzog and Woo [CHAN75] is very interesting for cases in which only one station (e.g., a CPU) in a queueing network model is to be analyzed under various system workload.

The parametric analysis of blocking queueing networks is executed as follows:

3.1. Computation of the Throughput Values of the Subsystem

3.2. Determining of the Composite Station Capacity

3.3. Analysis of the Two-Station Load Dependent Network

3.1. Throughput Analysis of Blocking Networks

We already have an algorithm to calculate the throughput of a blocking network with single server stations [AKYL85b,c]. Here we propose an extension of that algorithm for blocking networks with different station types.

The basic concept is that the state space of the blocking queueing network with K total number of jobs is transformed into the state space of a nonblocking queueing network with $\hat{K}$ total number of jobs. The number of states in both networks should be approximately the same, if not identical. This would indicate that Markov processes describing the evolution of both networks over time have an almost identical structure. That, in turn, would guarantee that the throughputs of both systems are almost equal.

The following steps are executed in order to compute the throughput values in queueing networks with blocking.

3.1.1. Determine the number of states in blocking queueing network.

3.1.2. Determine the total number of jobs $\hat{K}$ in the equivalent nonblocking queueing network.

3.1.3. Analyze the nonblocking queueing network with $\hat{K}$ jobs to obtain the throughput values which are equal to the throughput value of the blocking network with K jobs.

3.1.1. Determine the number of states in blocking queueing networks

As previously mentioned, in blocking networks each station has a capacity limit, which indicates that only a certain number of states can be feasible. The feasible states for blocking networks are obtained by realizing that the number of jobs in the i-th station k_i may not exceed its capacity M_i, $k_i \leq M_i$.

Blocking events which occur in networks with finite station capacities must also be taken into account. Therefore, the m_i (number of servers) neighbors of the feasible states are included, representing the blocking states. Whenever a transition occurs from one state to another state in which the capacity limit of a station would be violated, we assume that the transition causes a blocking action in the network and that the state entered is a blocking state. In reality, the job still resides in the source station. All the other states are infeasible and are cancelled.

Using this method we obtain a sub-state space for the blocking network. From the reduced state space we can obtain the number of states Z' of the blocking network, which is the sum of feasible states and blocking states. Since the number of states Z' can be very large for general networks we cannot draw the state space, eliminate the nonfeasible states or count the total number of feasible states and their neighbors as blocking states in an efficient way. In order to directly obtain the number of states Z' in queueing networks with blocking, we have developed an efficient convolution algorithm [AKYL85b,c] which is applicable only to networks containing Type 1a stations. We must find an efficient algorithm which provides the number of states Z' (the feasible states and their neighbors as blocking states) in closed queueing networks with four different finite capacity station types.

3.1.2. Determine the total number of jobs $\hat{K}$ in the equivalent nonblocking network

Our primary objective is to find an equivalent nonblocking network which has the same number of states and the same state space structure as the blocking network. In general, the state space of the blocking network cannot be transformed bijectively into the state space of an equivalent nonblocking network. However, the number of states in both systems may be equal or almost equivalent. This would imply that both systems have the same behavior and the throughputs of both systems are almost identical. Assume that the number of states Z' (feasible and blocking states) in the blocking queueing network is obtained somehow. The goal is to find a total number of jobs $\hat{K}$ in an equivalent network with infinite station capacities which will provide almost the same number of states, $\hat{Z}$, as in the blocking network. We then find an appropriate total number of jobs $\hat{K}$ in the equivalent nonblocking network such that $\hat{Z}$ will be approximately equal to Z'.

Since the number of states in both systems will be equal or almost equal, it implies that the Markov processes describing the evolution of both networks have approximately the same behavior. Consequently, the throughput of the equivalent nonblocking network $\lambda_{NB}(\hat{K})$ is almost equivalent to the throughput of the blocking network $\lambda_B(K)$.

3.1.3. Determine the throughput of the equivalent nonblocking network

By analyzing the equivalent nonblocking network with $\hat{K}$ total jobs using a product form algorithm such as mean value analysis, [REIS80], we obtain the total throughput $\lambda_{NB}(\hat{K})$. This is almost identical to the total throughput $\lambda_B(K)$ of the blocking network with K total number of jobs [AKYL85b].

$$\lambda_B(K) \; = \lambda_{NB}(\hat{K}) \tag{3}$$

Note that we do not need any other performance measures than the throughput of the subsystem. By varying the number of jobs in the subsystem, from 1 to K, we can obtain $\lambda_B(1), \lambda_B(2), \cdots, \lambda_B(K)$. Note that the throughput values $\lambda_B(k)$ for $k = 1,..,K$ for the blocking network are approximate. Consequently the parametric analysis for blocking networks will provide approximate results.

3.2. Determining the Capacity of the Composite Station

The given system has been reduced to a two-station blocking network. The service rate of composite station $\mu_c(k)$ is load-dependent, and set equal to the computed throughputs $\lambda_B(k)$ for $k = 1,...,K$. The major problem which arises is the capacity of the composite station. Initially, we define the capacity of the composite station as:

$$M_c = \sum_{j \in \sigma} M_j \tag{4}$$

where σ represents the subsystem.

Our experience shows that this overestimates throughput, since the shorted station can be blocked in the actual network with less than M_j jobs in the subsystem σ.

Another possibility is using a capacity weighted by the transition probabilities from the shorted station i to the stations in the subsystem:

$$M_c = \sum_{j \in \sigma} M_j \ p_{ij} \tag{5}$$

However, we have discovered that this underestimates the throughput.

Note that Suri/Diehl [SURI86] realized also this fact and solved this problem by constructing so-called "views" as seen by the previous station. They assume that the shorted station i must have infinite capacity. The "view" seen by station i is for k jobs in the successor stations sometimes sees the station i as blocked and sometimes unblocked. Thus, the shorted station i sees a finite buffer of variable size k in the composite station. They introduce a so-called variable buffer size model which represents the view seen by the shorted station i. An iterative formula is used to compute the views. There are some limitations in their work. The model is restricted to single server tandem networks, in particular, the job flow can only be in one direction. The network cannot have arbitrarily switched stations. Another limitation is that the shorted station must have an infinite capacity. Their work will definitely help us in our investigations.

3.3. Analysis of the Two-Station Load Dependent Network

In [AKYL85a] we have shown that two-station closed queueing networks with blocking have an exact product form solution. The solution concept is based on the transformation of the state space of the blocking queueing network into a state space for a nonblocking network. The state space of two-station queueing networks is one-dimensional. It is easy to find an equivalent nonblocking network which has exactly the same structure as the blocking network. In [AKYL85a,b] we have proved the following theorem:

Theorem. **For a two-station closed queueing network with classical blocking there exists an equivalent two-station closed queueing network without blocking having the same structure.** The equilibrium state probabilities p(k_1, k_2)

for the blocking network are computed by the product form solution for the equivalent nonblocking network.

In the load-dependent case the following problem can cause the results for two-station network to be only approximate. State transitions of the blocking and equivalent nonblocking network do not agree. This will be explained by a numerical example. Assume a two-station network with $K = 6$ jobs are given. The first station has the capacity $M_1 = 4$ and the second station $M_2 = 3$. The service rate of the first station is load-independent μ_1 where the second station has load-dependent service rates $\mu_2(k)$. The state space diagram for the above network is:

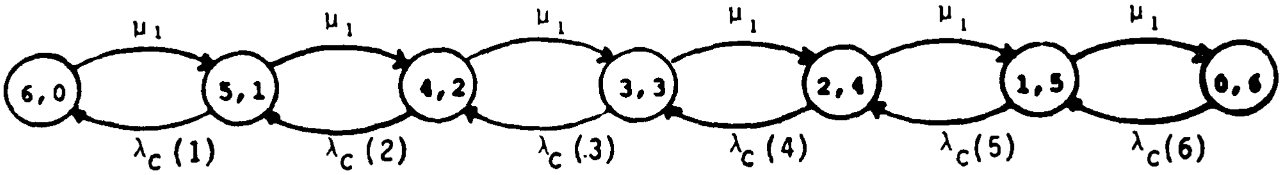


Figure 4.

By considering the station capacities the following sub-state space is obtained shown, in Figure 5, containing the feasible states and the blocking states (denoted by *) for the blocking network.

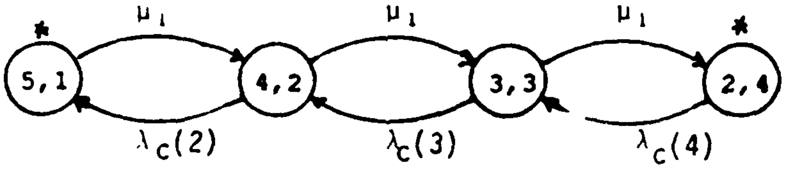


Figure 5.

Now we find a nonblocking network with an appropriate number of jobs $\hat{K}$ which provides the same state space structure as the blocking network. Since there are $Z' = 4$ states in Figure 5, we find that $\hat{K} = 3$. The state space for $\hat{K} = 3$ jobs is shown in Figure 6.

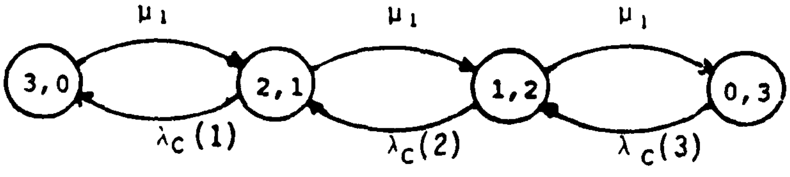


Figure 6.

However, the transition rates between states in Figure 6 and 5 do not agree. This can cause the results to no longer be exact. We can attempt to overcome this problem by having the same transitions between the states of the nonblocking network as shown in Figure 7.

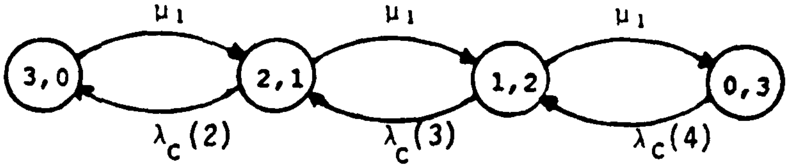


Figure 7.

Figure 7 and Figure 5 have exactly the same structure and the same transition rates. On the other hand, since we have a network with load-dependent service rates, transition rates of Figure 6 are more realistic than those of Figure 7. We will investigate further on which diagram should be analyzed, and explore the possibility of extending our exact algorithm [AKYL85a] to handle the "load-dependent" case.

4. Extended Parametric Analysis of Blocking Queueing Networks (PROPOSED CONCEPT II)

Some papers extending Norton's Theorem for queueing networks with infinite capacity have been published in the recent years. Kritzinger/van Wijk/Krzesinski [KRIT82] have extended the work of [CHAN75] to closed, open and mixed queueing networks with multiple job classes. Balsamo/Iazeolla [BALS82] partition a network with N stations and K jobs into two subsystems where the first subsystem contains the stations whose behavior is to be studied and the second subsystem represents the uninteresting part, i.e., stations whose behavior is not of interest. Their method is based on the matrix of the transition probabilities. They eliminate one uninteresting station by setting its mean service time equal to zero, and obtain a new transition probability matrix for the jobs. They repeat this elimination procedure for the next uninteresting station and construct a new transition probability matrix for the jobs. This elimination procedure must be repeated and a new transition probability matrix must be constructed for each station as they want to eliminate. This method is complex and requires a large amount of computation time for large queueing networks.

We will extend the parametric analysis of blocking queueing networks as proposed in section 3, in which a queueing network model can arbitrarily be partitioned into S disjoint subnetworks, in short form SNW_j (for $j = 1,2,..,S$), each containing multiple stations. This concept is illustrated by the example given in Figure 1. Figure 8 shows the given queueing network from Figure 1 decomposed into 4 subnetworks.

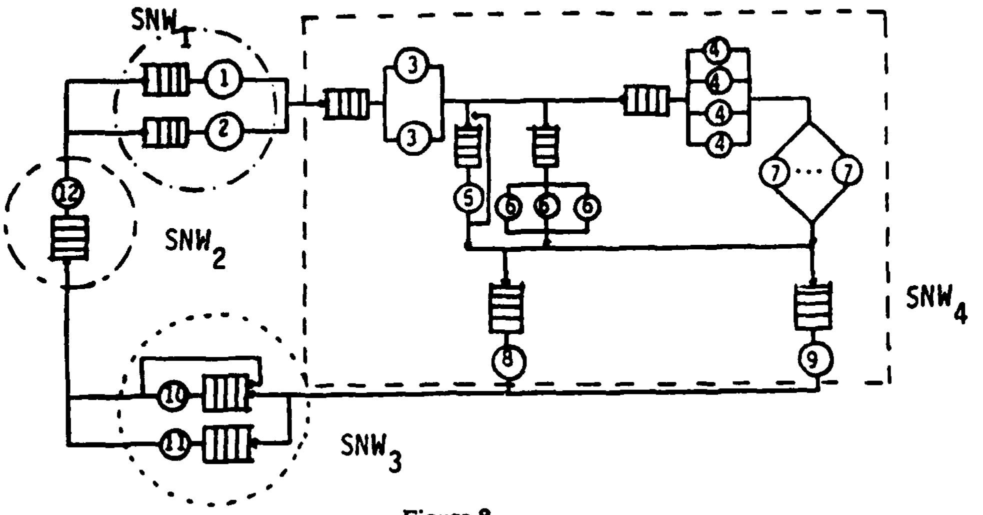


Figure 8.

Each subnetwork SNW_j is analyzed by shorting all other stations in all other subnetworks, i.e., their service times are set to zero. Since the subnetworks can be analyzed independently, this analysis can be executed in parallel. As a result, the load dependent throughput values $\lambda(k)$ for each subnetwork are obtained simultaneously. Each subnetwork SNW_j containing multiple stations can thus be composed into a single station. The computed load dependent throughputs $\lambda(k)$ for each SNW_j (for $1 < j < S$) are set equal to the load dependent service rates $\mu_c(k)$ of the respective composite station. The composite stations are serially switched and the simplified network, shown in Figure 10, is easily analyzed when computing all relevant characteristic performance measures. These measures are valid for each station of the originally given network model, Figure 1.

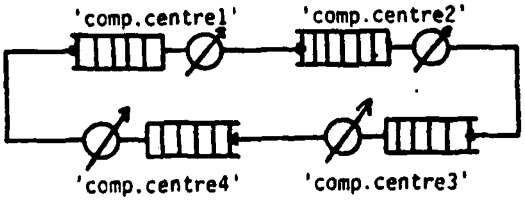


Figure 9.

This extended parametric analysis is not only motivated only by the desire to accelarate the processing speed and reduce the memory space, but also by the fact that some stations (more than one) could be studied under various system input parameters in which the remaining subsystem is represented by a composite station containing all stations whose behavior does not interest us.

As a formal example, assume that stations 1 and 2 in Figure 1 represent two independent CPU's. If we wish to investigate only these CPU's under various workload parameters, we do not need to consider the remaining stations (3 through 12) in our computations. It is sufficient to initially analyze the remaining system computing the composite load dependent throughput values $\lambda(k)$ by setting the mean service times of both CPU's (stations 1 and 2) to zero. In this way we obtain Figure 10, in which the composite station represents stations 3 through 12.

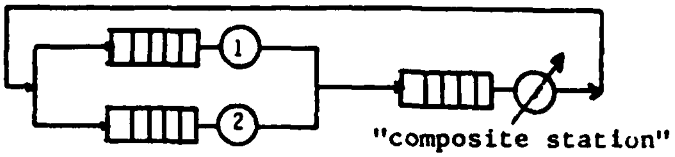


Figure 10.

The mean service times of the CPU's (stations 1 and 2) are stated initially , while the mean service rate of the composite station is equal to the computed load dependent throughput value $\lambda(k)$. This queueing network model, Figure 10, can be used for the analysis of the stations 1 and 2 under various system input parameters.

The extended parametric analysis is realized in three steps:

4.1. Computation of the load dependent throughputs

4.2. Composing of Subnetworks

4.3. Analysis of Serial Order

4.1. Computation of the load dependent throughputs for the subnetworks

In order to analyze a subnetwork SNW_j, we must short all stations which do not belong to that subnetwork SNW_j. We then compute the load dependent throughputs for each subnetwork SNW_j for $(1 < j < S)$. For the computation of this measure we will use the algorithm proposed in section 3.1. Note that each subnetwork can be analyzed independently from the others. The independent analysis of each subnetwork can be executed in parallel in order to accelarate the processing speed. In the infinite capacity network case we realized this parallel execution on 4 processors and reached an acceleration of the computations by a factor of 3 to 3.5 [AKYL86b]. An optimal decomposition aggregates the entire network into multiple subnetworks such that the following relation is valid:

$$\{Number\ of\ Subnetworks\}\ mod\ n \ \geq (n-1) \tag{6}$$

where n is the number of processors.

The purpose of this relation is to prevent processors from waiting on the results of other processors.

4.2. Composing of Subnetworks

We compose the stations of each subnetwork into one station. The mean service rate $\mu_c(k)$ of this composite station is dependent on the number of jobs in the subnetwork and is given by the throughput of the composite subnetwork.

$$\mu_{c_j}(k) = \lambda_{SNW_j}(k) \qquad \text{for } j = 1,...,S. \tag{7}$$

These composite stations are switched serially in an arbitrary order.

The same problem occurs here arises with the capacity of each composite station. The solution of section (3.2) will provide an answer to this problem.

4.3. Analysis of Serial Order

For the analysis of these arbitrarily ordered serially switched composite stations we apply the algorithm suggested in [AKYL85c]. Note that the algorithm suggested in [AKYL85c] is applicable to Type 1a stations. However, it can easily be extended such that other types of stations can also be analyzed.

REMARK

The Extended Parametric Analysis concept can be applied in order to simplify the computational requirements involved in large queueing network models with blocking. Using this concept, the large storage requirement and the long run times of the existing algorithms, in particular of our algorithm [AKYL85c] can be reduced drastically. This is due to the fact that the one large queueing network is analyzed as multiple small independent networks.

The major advantage of this technique is that computational expenses are reduced if only a few stations from the queueing network model are to be investigated under various system workloads. In this case we must determine the throughput values of the subnetwork which contains the interesting stations. These throughput results are then used for the analysis of the remaining subnetworks. The advantage results from the fact that the throughput values for the remaining subnetworks are computed only once in the beginning and remain fixed under various system workloads.

5. Validation

Our solutions in both cases (Parametric and Extended Parametric Analysis) will be approximate. For validation all suggested formulas and the proposed algorithms must be tested and implemented. A large variety numerical examples, including several stress tests, must be executed and then compared with simulation results. The RESQ [SAUE81] simulation package is used to simulate the blocking networks.

References

AKYL85a I. F. Akyildiz, "Exact Product Form Solution for Queueing Networks with Blocking", *to appear in IEEE Transactions on Computers.*

AKYL85b I. F. Akyildiz "On the Exact and Approximate Throughput Analysis of Closed Queueing Networks with Blocking", *to appear in IEEE Transactions on Software Engineering.*

AKYL85c I. F. Akyildiz, "Analysis of Closed Queueing Networks with Blocking", *Technical Report of Louisiana State University,* TR-85-046, September 1985.

AKYL85d I. F. Akyildiz, "Mean Value Analysis for Blocking Queueing Networks", *to appear in IEEE Transactions on Software Engineering.*

AKYL86a I. F. Akyildiz and S. Kundu, "Buffer Allocation in Deadlock Free Queueing Networks with Blocking", Technical Report of Louisiana State University, Baton Rouge, LA, TR-86-13, May 1986.

AKYL86b I. F. Akyildiz, G. Bolch and M. Paterok, "Parallel Processing of Queueing Network Models for Computer Systems," *Proc. of the Int. Conference on Simulation and Modeling, Naples/Italy,* Sept. 29 - Oct. 1, 1986.

BALS82 S. Balsamo and G. Iazeolla, "An Extension of Norton's Theorem for Queueing Networks", *IEEE Transactions on Software Eng., Vol. SE-8, No. 4, July 1982,* pp. 298-305.

BALS83 S. Balsamo and G. Iazeolla, "Some Equivalence Properties for Queueing Networks with and without Blocking", *Proceedings of Performance 83 Conference,* editors A. K. Agrawala and S. Tripathi, North Holland Publ. Co., 1983, pp. 351-360.

BCMP75 F. Baskett, K. M. Chandy, R. R. Muntz and G. Palacios, "Open, Closed and Mixed Network of Queues with Different Classes of Customers", *Journal of the ACM,* Vol. 22, Nr. 2, Apr. 1975, pp.248-260.

BOXM81 O. I. Boxma and A. G. Konheim, "Approximate Analysis Exponential Queueing Systems with Blocking," *Acta Informatica,* vol. 15, January 1981, pp. 19-66.

BUZE71 J. P. Buzen "Queueing Network Models of Multiprogramming" *PhD Thesis, Div. Eng. and Applied Sciences, Harvard Univ., Cambridge,* Mass. Aug. 1971.

CHAN75 K. M. Chandy, U. Herzog and L. Woo, "Parametric Analysis of Queueing Network Models, *IBM Journal Res. Dev.,* Vol. 19, Nr. 1, Jan. 1975, pp.43-49.

GORD67a W. J. Gordon and G. F. Newell, "Closed Queueing Systems with Exponential Servers", *Operations Research,* 15, 1967, pp. 254-265.

GORD67b W. J. Gordon and G. F. Newell, "Cyclic Queueing Systems with Restricted Queues", *Operations Research,* 15, Nr. 2, April 1967, pp. 266-277.

HORD81 A. Hordijk and N. van Dijk, "Networks of Queues with Blocking," *Proceedings,* 8th International Symposium on Computer Peformance Modelling, Measurement, and Evaluation, Amsterdam, November 4-6, 1981.

JACK63 J. J. Jackson, "Jobshop-like Queueing Systems", *Management Science,* 10, 1, 1963, pp. 131-142.

KONH76 A. G. Konheim and M. Reiser, "A Queueing Model with Finite Waiting Room and Blocking," *Journal of the ACM,* vol. 23, Number 2, April 1976, pp. 328-341.

KONH77 A. G. Konheim and M. Reiser, "Finite Capacity Queueing Systems with Applications in Computer Modeling", *SIAM Journal on Computing*, Vol.7, No. 2, May 1977, pp. 210-229.

KRIT82 P. S. Kritzinger, S. van Wyk and A. E. Krzesinski, "A Generalization of Norton's Theorem for Multiclass Queueing Networks", *Performance Evaluation, Vol. 2, Oct. 1982, pp. 98-107.*

PERR81 H. G. Perros, "A Symmetrical Exponential Open Queue Network with Blocking and Feedback", *IEEE Transactions on Software Engineering*, SE-7, 1981,pp. 395-402.

PERR84 H. G. Perros, "Queueing Networks with Blocking: A Bibliography", *ACM Sigmetrics* Performance Evaluation Review, August 1984.

PERR86 H. G. Perros and T. Altiok, "Approximate Analysis of Open Networks of Queues with Blocking: Tandem Configurations", *IEEE Transactions on Software Engineering*, Vol. SE-12, No.3, March 1986, pp. 450-462.

PITT79 B. Pittel, "Closed Exponential Networks of Queues with Saturation: The Jackson Type Stationary Distribution and Its Asymptotic Analysis," *Mathematics of Operations Research*, vol. 4, 1979, pp. 367-378.

REIS80 M. Reiser and S. S. Lavenberg, "Mean Value Analysis of Closed Multichain Queueing Networks", *Journal of the ACM, Vol. 27, No. 2, April 1980, pp. 313-322.*

SAUE81 C. H. Sauer and K. M. Chandy, "Computer Systems Performance Modeling" *Prentice Hall*, Englewood Cliffs, N. J., 1981.

SURI84 R. Suri and G. W. Diehl, "A new building block for performance evaluation of queueing networks with finite buffers", *ACM Sigmetrics Conference Proceedings*, Cambridge, Mass., Aug. 1984, pp.134-142.

SURI86 R. Suri and G. W. Diehl, "A Variable Buffer-Size Model and its Use in Analyzing Closed Queueing Networks with Blocking", *Management Science*, Vol. 32, No. 2, February 1986, pp. 206-225.

TAKA80 Y. Takahashi, H. Miyahara and T. Hasegawa, "An approximation Method for Open Restricted Queueing Networks", *Operations Research*, Vol. 28, Nr. 3, May-June 1980, pp. 594-602